(19) KOREAN INTELLECTUAL PROPERTY OFFICE

KOREAN PATENT ABSTRACTS

(11)Publication number: 100202292 B1
(43)Date of publication of application: 19.03.1999

(21)Application number: 1019960065620
(22)Date of filing: 14.12.1996

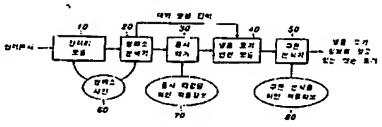
(71)Applicant: KOREA ADVANCED INSTITUTE OF SCIENCE AND TECHNOLOGY
(72)Inventor: LEE, SANG HO
OH, YEONG HWAN

(51)Int. Cl G06F 17/20

(54) DOCUMENT ANALYZER FOR CHINESE WRITING TEXT-TO-SPEECH SYSTEM

(57) Abstract:

PURPOSE: A document analyzer for a Chinese writing text-to -speech system is provided to offer a text-to-speech reliably by deciding a pronunciation display of a paragraph, a part-of-speech in morphemic and a dependent tree using a statistical language processing method, thereby offering an accurate pronunciation display and a syntax structure in a cadence creating mode deciding nature character of composite sound.



CONSTITUTION: The full treatment module(10) changes a character which is not a Korean to a Korean character in input document for changing voice conversion using a nondeterministic finite automata in a document analyzer and offers each sentence to a morphemic analyzer(20). The morphemic analyzer(20) gains all and available analyzing results and non-register paragraph is assumed when the analyzing of morphemic is fail, and reduces a non-register paragraph for analyzing. A part-of-speech tagger(30) extracts an optimum morphemic row in input sentence applied through morphemic analyzing based on probability information received from speech group. A pronunciation display conversion module(40) gains a pronunciation display of each paragraph using the result of a morphemic analyzing row applied from the part-of-speech tagger(30).

Legal Status
Date of final disposal of an application (19990223)
Patent registration number (1002022920000)
Date of registration (19990319)
Number of opposition against the grant of a patent ( )
Date of opposition against the grant of a patent ( )

Number of trial against decision to refuse ( )
Date of requesting trial against decision to refuse ( )
Date of extinction of right ( )

(19) 대한민국특허청(KR)
(12) 등록특허공보(B1)

(51) Int. Cl.[6]
G06F 17/20

(45) 공고일자 1999년06월15일
(11) 등록번호 10-0202292
(24) 등록일자 1999년03월19일

(21) 출원번호 10-1996-0065620
(22) 출원일자 1996년12월14일
(65) 공개번호 특1998-0047177
(43) 공개일자 1998년09월15일

(73) 특허권자 한국과학기술원 윤덕용
대전광역시 유성구 구성동 373-1
(72) 발명자 오영환
대전광역시 유성구 신성동 161-1 한울아파트 107동 404호
이상호
충청남도 서산시 동문동 311-7
(74) 대리인 김원호, 송만호

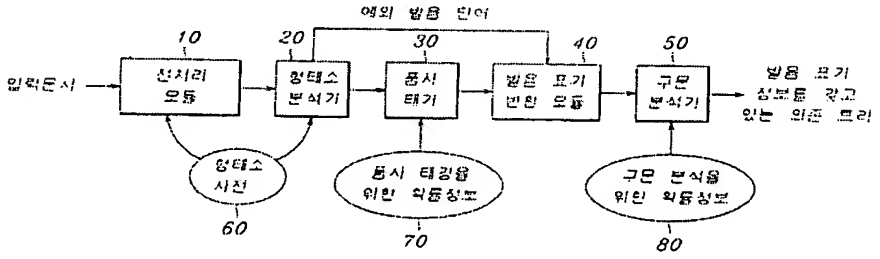
심사관 : 이은철

(54) 한문어 문서 음성 변환 시스템을 위한 문서 분석기

요약

주어진 입력 문서를 음성으로 변환시키는 문서 음성 변환 시스템에서 통계적 언어 처리 기법을 도입하여 합성음의 자연성을 좌우하는 운율 생성 모듈에 더욱 정확한 발음 표기 및 구문 구조를 제공하도록 한 것으로, 문서 음성 변환을 위하여 입력되는 문서를 하나의 문장씩 추출하며, 비결정 유한 오토마타를 이용하여 비한글 문자들을 한글 문자로 변환시키는 전처리 수단과, 상기 전처리 수단에서 인간되는 하나씩의 문장 어절에서 모든 가능한 형태소 분석 결과를 구하는 형태소 분석수단과, 상기 형태소 분석되어 인가되는 입력 문장을 확률 정보를 기반으로 하여 가장 가능성이 높은 형태소 분석 결과를 선택하여 최적 형태소 분석열을 추출하는 품사 태거와, 상기 품사 태거에서 인가되는 형태소 분석열의 결과를 이용하여 각 어절들의 발음 표기를 구하는 발음 표기 변환수단 및, 확률 의존 문법을 이용하여 어절들의 지배소-의존소 관계를 구하며, 입력된 문서에 대한 어절의 발음 표기 및 형태소 품사열, 의존 트리를 최종 결과로 출력하는 구문 분석수단을 구비하여 품사 태거에서 85.11%의 정확률을, 구문 분석기에서 78.68%의 정확률을 각각 나타내며, 특히 품사 태거의 경우, 처리 중인 문서에 미등록어가 없을 경우에는 96.46%의 높은 정확률을 제공한다.

대표도



명세서

도면의 간단한 설명

제1도는 본 발명에 따른 한국어 문서 음성 변환 시스템을 위한 문서 분석기 구성 블록도이고,

제2도는 본 발명에서 전처리를 위한 오토마타(automata)의 예를도이다.

제3도는 제1도의 본 발명에서 형태소 분석기의 구성도이고,'

제4도는 본 발명에서 나는에 대한 형태소 격자 구성이며,

제5도는 본 발명에서 격자 구성에 대한 알고리즘이다.

제6도는 본 발명에서 신을 신고 신고하기의 형태소 분석 격자 구성이고,

제7도는 본 발명에서 신을 신고 신고한다.의 의존 트리 구성이며,

제8도는 본 발명에서 신을 신고 신고한다.의 구구조 트리 구성이다.

발명의 상세한 설명

발명의 목적

발명이 속하는 기술 및 그 분야의 종래기술

본 발명은 주어진 입력 문서를 음성으로 변환시키는 문서 음성 변환 시스템(Text-to-Speech System)에 관한 것으로, 보다 더 상세하게는 통계적 언어 처리 기법을 도입하여 합성음의 자연성을 좌우하는 운율 생성 모듈에 더욱 정확한 발음 표기 및 구문 구조를 제공하도록 한 한국어 문서 음성 변환 시스템을 위한 문서 분석기에 관한 것이다.

일반적으로 문서 음성 변환 시스템은 주어진 입력 문서를 음성으로 변환하는 시스템으로, 맹인용 독서기 등 많은 응용 분야를 갖고 있다.

이러한 문서 음성 변환 시스템은 크게 문서 분석기와 운율 생성 및 신호 합성 모듈로 이루어지며, 문서 분석기는 단어의 발음 표기와 그 문장이 갖고 있는 운율이 단어의 품사와 문장의 구문 구조가 밀접한 관계를 갖고 있기 때문에 이를 분석하는데 사용된다.

기존의 한국어 문서 분석기들은 형태소 분석 단계에서 최장 일치법 등을 이용하여 하난의 결과만을 추출하고, 이를 이용하여 발음 변환, 구문 분석 등을 수행한다.

이외에 최근에는 두 단계(two-level) 모델에 기반한 발음 표기 변환 모듈이 개발되었으며, 더욱 정확한 정보를 추출하기 위해 많은 연구가 진행중이다.

발명이 이루고자하는 기술적 과제

그러나 전술한 바와 같은 문서 음성 변환 시스템은 프로그래머가 설정한 의미의 품사로써 입력되는 문서의 품사 정보를 분석하게 되므로 같은 음절이 서로 다른 품사를 갖게 되는 경우 품사 정보의 그릇된 판단으로 발음 표기에 오 변환을 일으키게 되는 문제점이 있어 문서 음성 변환에 신뢰성이 저하되는 문제점이 있었다.

본 발명은 이와 같은 문제점을 감안하여 안출한 것으로, 그 목적은 통계적 언어 처리 기법을 이용하여 어절의 발음 표기와 형태소 품사열 및 의존 트리의 결정으로 합성음의 자연성을 좌우하는 운율 생성 모듈에 더욱 정확한 발음 표기 및 구문 구조를 제공하여 신뢰성 있는 문서 음성 변환을 제공하도록 한 것이다.

이와 같은 목적을 달성하기 위한 본 발명은 음성 변환을 위하여 이력되는 문서를 하나의 문장식 추출하며, 비결정 유한 오토마타를 이용하여 비한글 문자들을 한글 문자로 변환시키는 전처리 수단과;

상기 전처리 수단에서 인가되는 하나씩의 문장 어절에서 모든 가능한 형태소 분석 결과를 구하는 형태소 분석수단과;

상기 형태소 분석되어 인가되는 입력 문장을 확률 정보를 기반으로 하여 가장 가능성이 높은 형태소 분석 결과를 선택하여 최적 형태소 분석열을 추출하는 품사 태거와;

상기 품사 태거에서 인가되는 형태소 분석열의 결과를 이용하여 각 어절들의 발음 표기를 구하는 발음 표기 변환수단 및;

확률 의존 문법을 이용하여 어절들의 지배소-의존소 관계를 구하며, 입력된 문서에 대한 어절의 발음 표기 및 형태소 품사열, 의존 트리를 최종 결과로 출력하는 구문 분석수단을 구비하는 것을 특징으로 한다.

발명의 구성 및 작용

이하, 첨부된 도면을 참조하여 본 발명의 바람직한 일 실시예를 상세히 설명하면 다음과 같다.

제1도에서 알 수 있는 바와 같이 본 발명에 따른 한국어 문서 음성 변환 시스템을 위한 문서 분석기에서 전처리 모듈(10)은 음성 변환을 위하여 입력되는 문서를 비결정 유한 오토마타(nondeterministic finite automata)를 이용하여 비한글 문자들을 한글 문자로 바꾸고 형태소 분석기(20)측에 문장 하나씩을 제공한다.

형태소 분석기(20)는 모든 가능한 형태소 분석 결과를 얻고, 형태소 분석이 실패되었을 경우는 분석중인 어절에 미등록어가 있다고 판단하고 미등록어를 추정한 후, 언어적 휴리스틱을 이용하여 분석을 위한 미등록어 후보들의 수를 줄인다.

품사 태거(part-of-speech tagger :30)에서는 말뭉치로부터 얻은 확률 정보를 기반으로 하여 형태소 분석을 통해 인가되는 입력 문장에서 최적 형태소 분석열을 추출한다.

발음 표기 변환 모듈(40)에서는 품삼 태거(30)에서 인가되는 형태소 분석열의 결과를 이용하여 각 어절들의 발음 표기를 구한다.

구문 분석기(50)는 확률 의존 문법을 이용하여 어절들의 지배소-의존소 관계를 구하며, 입력된 문서에 대한 어절의 발음 표기 및 형태소 품사열, 의존 트리를 최종 결과로 출력한다.

또한, 품사 태거(30)와 구문 분석기(50)에는 중의성 해소를 수행하기 위해 각각 품사 태깅을 위한 확률 정보(70)와 구분 분석을 위한 확률 정보(80)를 구비한다.

전술한 바와 같은 기능을 구비하여 이루어지는 본 발명에서 입력되는 한국어 문서의 음성 변환 동작을 설명하면 다음과 같다.

음성 변환을 위한 문서가 전처리 모듈(10) 입력되면 총 48개의 상태로 이루어진 비결정 유한 오토마타가 이용되어, 이 중 12개의 종결 상태(final state)가 각각 다른 비한글의 한글 변환 모듈로 구현되어 있는 전처리 모듈(10)은 입력 문서로부터 하나씩 문장을 추출함과 동시에 그 문장안에 있는 비한글 문자들을 모두 한글로 변환시킨다.

이때, 비한글 문자들은 영어, 영어 약어, 숫자, 전화번호, 년도, 시간 등으로 구별되고 이들은 다음에 오는 한글 문자들을 고려하여 한글화된다.

일 예를들어 Mr. Lee의 전화번호 이라는 문서가 입력되는 경우 첨부된 제2도의 전처리를 위한 오토마타에서 알 수 있는 바와 같이, 'Mr.'에 의해 4 번 상태에 도달하게 되고, 'Mr.'가 사전에 등록되어 있으면 '미스터'로 한글화되고, 그렇지 않을 경우에는 '엠 알'로 한글화된다.

이때, 'Mr.'는 사전에 등록되어 있는 영어 약어이므로 '미스터'로 한글화된 그 다음 'Lee'의 경우는 5번 상태에서 'Lee'나 '리'로 한글화되고, 방금전에 입력된 '의'를 다시 입력하여 2번 상태에서 한글을 처리한다.

이와 같은 과정으로 미스터 리의 전화번호는이라는 최종 결과를 얻게 된다.

이후, 형태소 분석기(20)는 전술한 바와 같은 동작을 통해 전처리가 완료되어 인가되는 문장 어절의 오른쪽에서 왼쪽으로 형태소들을 찾아 형태소 격자를 구성하게 되는데, 첨부딘 제3도에서 알 수 있는 바와 같이 입력 어절(21)에 대하여 위치 추정모듈(22)을 통해 불규칙 및 축약 현상 위치들을 미리 계산하고, 격자 구성모듈(23)에 설정되어 있는 알고리즘에 의해 형태소 분석을 수행한다.

형태소 분석 결과는 형태소 격자로 표현하는데, 예를 들면, 나는의 최종 분석 결과는 첨부된 제4도와 같고, 'INI'부터 'FIN'까지의 모든 가능한 경로가 각각 서로 다른 분석 결과를 나타낸다.

이때, 첨부된 제4도와 같은 형태소 격자를 표현하기 위해서 하기의 [표 1]에 보이는 집합 L을 정의하였다.

집합 L의 원소(k,w,t,l)는 형태소 격자에서 하나의 노드와 그 노드의 오른쪽에 붙어 있는 에지들을 표현하며, 제4도에 해당하는 집합 L은 하기의 [표 2]와 같다.

[표 1]

$$L = \{ l \mid l = (k,\ w,\ t,\ l) \}$$

k : l의 인덱스

w : 형태소

t : 품사

l : l이 가리키고 있는 원소들의 인덱스 집합

[표 2]

L = { (0, FIN, NULL , {}),

(1, ㄴ, 전성어미 , {0}),

(2, 는, 전성어미 , {0}),

(3, 는, 주제격조사 , {0}),

(4, 날, 동사 , {2}),

(5, 나, 보조용언 , {2}),

(6, 니, 동사, , {2}),

(7, 나, 인칭대명사 , {3}),

(8, IN1, NULL , {7, 6, 5, 4}) }

한편, 격자 구성 모듈(23)에 수록되는 격자 구성 알고리즘은 우선 어절의 자소의 열로 만든 후, 알고리즘의 편의를 위해 각 자소 사이에 0부터 시작해서 지정되는 스텝 (step)만큼씩 증가하며 숫자를 삽입한다.

예를 들어, 스텝을 6으로 두었을 때 나는의 경우, '0 ㄴ 6 ㅏ 12 ㄴ 18 ㅡ24 ㄴ 30'이된다.

이와 같이 어절을 자소열로 표현한 후, 집합 L을 얻는 과정은 첨부된 제5도에 제시된 알고리즘을 사용하게 되는데, wi,j는 숫자 I와 j사이의 자소열을 뜻하고, 함수 LookupDict 는 사전으로부터 입력 자소열 wi,j의 가능한 품사들을 찾는 기능을 한다.

함수 SearchL 은 집합 L에서 (wi,j, t)가 접속할 수 있는 원소들의 인덱스 집합을 받는 기능을 하는 것으로 이는 미리 구축된 품사 접속표를 이용하여 이루어진다.

알고리즘에서 사용한 집합 J는 현재까지 완성된 격자에서 'FIN' 노드까지 경로가 존재하는 노드들의 왼쪽 번호들을 모아둔 집합으로, 이는 불필요한 사전 탐색을 줄이기 위해 사용되었다.

불규칙을 처리하는 부분은 미리 구해진 불규칙 위치 정보를 기반으로 이루어지는데, 나는의 경우, 12와 24에서 'ㄹ' 탈락 현상이 발생될 수 있으므로, j가 12 혹은 24일 때 'ㄹ'을 첨가하고 새로운 변수 h를 j부터 0까지 감소시키며, wh,j를 사전에서 찾게 된다.

그러므로 집합 L에 있는 (4, 날, 동사, 2)는 j가 12일 때 첨가된 것이다.

한편, 분석중인 어절이 미등록어를 포함하는 경우에는 미완성된 형태소 격자로부터 모든 가능한 미등록어 후보를 생성하여 격자를 완성하게 된다.

이 때, 조사나 어미와 같은 기능어들은 모두 사전에 등록되어 있다고 가정하면 체언 혹은 용언과 같은 내용어만이 미등록어가 될 수 있으므로, 미등록어는 항상 어절의 왼쪽 부분에 나타나게 된다.

그러므로 미등록어를 추측한다는 것은 왼쪽에 남아 있는 자소열을 오른쪽에 있는 노드의 품사와 접속 가능한 품사로 할당하는 것이 되며, 이러한 방법으로 미등록어를 추정한 형태소 격자를 생성하게 된다.

그러나, 미등록어가 추정된 형태소 격자는 가능한 형태소 분석 수가 너무 많게 되므로 다음 단계인 품사 태깅에 많은 오류를 범하게 할 수 있어 이를 방지하기 위해 음절 정보와 단서(clue) 형태소를 이용하여 후보의 수를 줄인다.

먼저, 음절 정보를 사용하는 방법은 미등록어의 마지막 음절이 추정된 품사의 마지막 음절로 사용되지 않을 경우에는 이를 제외하는 것이다.

예를 들면, 한국어의 용언 중 마지막 음절이 '느'인 단어는 총 6개 뿐이라는 사실을 이용하면 추정된 미등록어의 품사가 용언이고 그 미등록어의 마지막 음절이 '느'일 경우 그 노드를 격자에서 제외할 수 있게 된다.

단서 형태소를 이용하는 방법은 미완성인 형태소 격자내에 아주 빈도가 높고 그 어절의 구성을 추정하기에 충분하다고 생각되는 형태소가 발견되면, 그 형태소의 앞에 추정된 미등록어만 남기고 나머지를 격자

에서 제거한다.

예를 들면, 우회시켜라는 어절에서 '우회'가 미등록어일 경우, 미등록어 후보로는 '우회/동작성보통명사', '우회시키/동사, '우회시키/형용사' 등을 얻게 되는데, '시키'라는 형태소가 단서 형태소이므로 '우회/동작성보통명사'만을 남기고 나머지는 모두 제거하게 된다.

이와 같은 방법을 이용하여 미등록어가 포함된 어절에 대해 어절당 형태소 분석 개수를 22.01개에서 10.83개로 줄일 수 있게 된다.

이와 같이 입력되는 문장에 대하여 형태소 분석이 완료되면 품사 태거(30)는 각 어절들의 형태소 분석 후보들 중 최적의 형태소 분석 결과를 차기 위하여 미등록어를 처리하는 방법에 중점을 두어서 확률에 기반한 품사 태깅을 수행한다.

확률에 기반한 품사 태깅은 n개의 어절로 구성된 문장, 즉 어절열 w1w2w3...wn인 w1..n에 대해 최적의 형태소 분석 결과를 찾는 문제이므로, i 번째 어절의 형태소 분석 결과를 형태소열 mi와 품사열 ti의 쌍으로 표시하여 다음과 같이 품사 태깅 함수 ϕ(w1..n)를 정의한다.

$$\phi(w_{1..n}) \cong \arg\max_{m_{1..n},t_{1..n}} P(m_{1..n},t_{1..n}|w_{1..n}) \quad (1)$$

식 1을 베이즈 룰과 일차 마르코프 가정 등을 이용하면 다음과 같은 식으로 표현된다.

$$\phi(w_{1..n}) \cong \arg\max_{m_{1..n},t_{1..n}} \prod_{i=1}^{n} P(m_i|t_i)P(t_i|t_{i-1}) \quad (2)$$

식 2는 결국 이전 어절의 품사열에서 현재 어절의 품사열로 천이할 확률과 어절의 품사열에서 임의의 형태소열이 발생될 확률들을 매 어절마다 곱하였을 때 가장 큰 값을 갖는 형태소 분석열을 찾는 의미가 된다.

이것은 형태소 분석 결과를 노드로, 인접하는 어절들의 형태소 분석 결과 사이에 에지(edge)를 두어 노드에는 P(mi/ti)를, 에지에는 P(ti/ti-1)을 할당하고 가장 높은 확률을 내는 경로를 취하는 것으로 볼 수 있다.

예를 들면, 신을 신고 신고하기라는 어절열은 첨부된 제6도와 같이 나타낼 수 있고, 이런 종류의 문제는 바이터비(Viterbi) 알고리즘에 의해 구해진다.

이때, nc : 보통명사, po : 목적격 조사, vb : 동사, ex : 보조적 연결어미, ec : 연결어미, xj : 형용사 파생 접미사, en : 명사형 정성어미, na : 동작성 보통명사, xv : 동사 파생 접미사, vx : 보조 용언으로 정의한다.

그러나, 식 2를 그대로 이용하기에는 자료의 부족 현상이 발생할 수 있으므로 이를 더 작은 단위들의 확률 값으로 근사하여 모델의 파라미터 수를 줄인다.

우선, 식 2의 P(ti/ti-1)에서 ti는 형태소 품사열을 뜻하므로, 이를 $t_i^1 \cdots t_i^{N_i}$ 로 풀어 쓸 수 있다.

여기서 t_i^j는 i번째 어절의 임의의 형태소 분석 결과에서 j번째 형태소 품사를 뜻하는 것이고 Ni 는 그 형태소 분석 결과에 사용된 품사의 갯수를 뜻한다.

예를 들면, t_i가 vb, ex, vx, en 이라면 t_i^3 은 vx이고, N_i는 4가 된다.

이때, vb : 동사, ex : 보조적 연결어미, vx : 보조 용언, en : 명사형 전성어미로 정의한다.

이러한 방법을 사용하면 P($t_i|t_{i-1}$)은 다음과 같이 전개될 수 있다.

$$P(t_i|t_{i-1}) \equiv P(t_i^1 \cdots t_i^{N_i}|t_{i-1}^1 \cdots t_{i-1}^{N_{i-1}}) \quad (3)$$

$$\cong P(t_i^1 \cdots t_i^{N_i}|t_{i-1}^{N_{i-1}}) \quad (4)$$

$$\cong P(t_i^1|t_{i-1}^{N_{i-1}})P(t_i^2|t_i^1 t_{i-1}^{N_{i-1}})$$

$$\cdots\cdots P(t_i^{N_i}|t_i^{N_i-1} \cdots t_i^1 t_{i-1}^{N_{i-1}}) \quad (5)$$

$$\cong P(t_i^1, t_i^{N_i}) \prod_{j=2}^{N_i} P(t_i^j | t_i^{j-1}) \quad (6)$$

식 4는 식 3에서 현재 처리 중인 어절의 품사열은 이전 어절의 마지막 형태소 품사에만 의존한다는 가정에 의해 얻은 것이고, 이는 다시 연쇄 규칙(chain rule)에 의해 식 5로 전개되며, 일차 마르코프 가정을 적용하여 식 6을 얻게 된다.

즉, 품사열간의 천이 확률을 이전 어절의 마지막 형태소 품사에서 현재 처리중인 어절의 첫 형태소 품사로 천이하는 확률과 어절 내에서의 품사 천이 확률들을 곱한 것으로 근사한 것이다.

식 2의 P($m_i | t_i$)는 품사열에서 형태소열이 발생할 확률인데 여기서 고려해야 할 점은 첫 번째 형태소 m_i^1 이 미등록어일 수 있다는 점이다.

그러므로 m_i^1 이 등록어인지 미등록어인지를 뜻하는 새로운 변수 k_i^1 을 도입하여 다음과 같이 식을 다시 정의한다.

$$P(m_i | t_i) = P(m_i, k_i^1 | t_i) \quad (7)$$

식 7에서 k_i^1 은 m_i^1 이 등록어일 때 1을, 미등록어일 때 0을 취하게 하면, 다음과 같이 전개된다.

$$P(m_i, k_i^1 | t_i) \cong P(k_i^1 | t_i) P(m_i | t_i, k_i^1) \quad (8)$$

$$\cong P(k_i^1 | t_i) \cdot$$

$$[k_i^1 P(m_i | t_i, k_i^1 = 1) + (1 - k_i^1) P(m_i | t_i, k_i^1 = 0)] \quad (9)$$

식 9에서 $P(k_i^1 | t_i)$ 는 임의의 형태소 품사열이 주어지고, 그 중 첫 번째 품사가 미등록어인지, 혹은 등록어인지에 대한 확률을 뜻하고 이것은 독립 가정을 이용하면 다음과 같이 근사시킬 수 있다.

$$P(k_i^1 | t_i) \equiv P(k_i^1 | t_i \cdots t_i^{N_i}) \quad (10)$$

$$\cong P(k_i^1 | t_i^1) \quad (11)$$

한편, 식 9의 P(i/ti, k*)은 다음과 같이 전개할 수 있다.

$$P(m_i | t_i, k_i^1) \equiv P(m_i^1 \cdots m_i^{N_i} | t_i^1 \cdots t_i^{N_i} k_i^1) \quad (12)$$

$$= \frac{P(k_i^1 m_i^1 t_i^1 \cdots m_i^{N_i} t_i^{N_i})}{P(t_i^1 \cdots t_i^{N_i} k_i^1)} \quad (13)$$

$$= P(t_1^{N_1})P(m_1^{N_1}|t_1^{N_1})P(t_1^{N_1-1}|t_1^{N_1}m_1^{N_1})$$

$$\cdot P(m_1^{N_1-1}|t_1^{N_1-1}t_1^{N_1}m_1^{N_1})$$

$$\cdots P(k_1^1|t_1^2\cdots t_1^{N_1}m_1^2\cdots m_1^{N_1})$$

$$\cdot P(m_1^1|k_1^1t_1^1\cdots t_1^{N_1}m_1^2\cdots m_1^{N_1})$$

$$/P(t_1^1\cdots t_1^{N_1}k_1^1) \quad (14)$$

$$\cong P(m_1^1|k_1^1t_1^1\cdots t_1^{N_1}m_1^2\cdots m_1^{N_1})\cdot$$

$$\prod_{j=2}^{N_1}P(m_1^j|t_1^j) \quad (15)$$

식 15는 다음의 식 16, 17과 같은 가정을 이용하여 얻은 것으로, 최종적으로 k_1^1은 식 15의 첫 항에만 영향을 미치게 된다.

$$P(t_1^j|t_1^{j-1}\cdots t_1^{N_1}m_1^{j-1}\cdots m_1^{N_1}) \cong P(t_1^j|t_1^{j-1}\cdots t_1^{N_1}) \quad (16)$$

$$P(m_1^j|t_1^j\cdots t_1^{N_1}m_1^{j-1}\cdots m_1^{N_1}) \cong P(m_1^j|t_1^j) \quad (17)$$

식 15의 첫 항은 k_1^1의 값에 의해 다르게 처리되는데, k_1^1이 1일 경우는 식 18의 근사식을 이용하고, 0일 경우는 식 19와 같은 가정을 이용한다.

식 19는 한국어에서 미등록어에 대한 추정은 미등록어 다음에 오는 형태소와 품사에 의존한다고 가정한 것으로 일종의 언어적 유리스틱이라고 볼 수 있다.

$$P(m_1^1|k_1^1=1,t_1^1\cdots t_1^{N_1}m_1^2\cdots m_1^{N_1}) \cong P(m_1^1|t_1^1) \quad (18)$$

$$P(m_1^1|k_1^1=0,t_1^1\cdots t_1^{N_1}m_1^2\cdots m_1^{N_1}) \cong P(t_1^1|t_1^2m_1^2) \quad (19)$$

위의 두 식을 이용하게 되면, $P(m_{1,1}|k_1^1,t_{1,1})$는 k_1^1의 값에 따라 다음과 같은 근사 식을 얻게 된다.

$$P(m_{1,1}|k_1^1=1,t_{1,1}) \cong \prod_{j=1}^{N_1}P(m_1^j|t_1^j) \quad (20)$$

$$P(m_{1,1}|k_1^1=0,t_{1,1}) \cong P(t_1^1|t_1^2m_1^2)\prod_{j=2}^{N_1}P(m_1^j|t_1^j) \quad (21)$$

위의 두 식들과 식 6, 9를 식 2에 넣으면 다음과 같은 품사 태깅 수식을 얻을 수 있다.

$$\phi(w_{1\cdots n}) = \arg\max_{m_{1,1}\cdots m_{n,1}} \prod_{i=1}^{n}[P(m_{i,1}|t_{i,1})\prod_{j=2}^{N_i}P(t_i^j|t_i^{j-1})]P(t_i^1|t_{i-1}^{N_{i-1}}) \quad (22)$$

$$P(m_i|t_i) \cong P(k_i^1|t_i^1)[k_i^1 \prod_{j=1}^{N} P(m_i^j|t_i^j) +$$

$$(1-k_i^1)P(t_i^1|t_i^2 m_i^2) \prod_{j=2}^{N} P(m_i^j|t_i^j)] \qquad (23)$$

식 23에서 $P(k_i^1|t_i^1)$은 임의의 품사가 주어졌을 때 미등록어 혹은 등록어가 발생할 확률을 의미하는 것이고, 이 확률 값들은 미등록어가 포함된 말뭉치로부터 얻게 된다.

특히 이 값들은 현재 사용중인 사전의 표제어 수에 의존하여, 사전의 크기에 관계없이 확률적으로 최적의 품사 태깅 결과를 얻는다.

지금까지 전개한 식은 첨부된 제6도에서 노드에는 $P(m_i|t_i) \prod_{i=2}^{N} P(t_i^j|t_i^{j-1})$을, 에지에는 $P(t_i|t_{i-1})$을 각각 할당한 후, 바이터비 알고리즘으로 해결할 수 있다.

한편, 제6도의 노드들을 관찰해 보면, 노드들이 형태소 분석기의 결과인 형태소 격자를 풀어놓은 것임을 알 수 있으므로 어절들의 형태소 격자를 연결하여 바이터비 알고리즘을 수행하면 결과를 더욱 빠르게 얻을 수 있다.

이와 같은 동작에 의하여 각 어절들의 형태소 분석이 완료되면 발음 표기 변환 모듈(40)은 주어진 문장의 발음 표기를 찾기 위해 문장을 이루는 모든 어절들의 쓰임새를 정확히 인식하기 위하여 품사 태거의 결과를 기반으로 하여 동일한 어절일지라도 쓰임새에 맞게 다르게 발음될 수 있도록 처리한다.

이때 사용된 발음 변환 규칙들은 문교부에서 고시한 '표준 발음법'을 적용한다.

발음 표기 변환을 하는 방법은, 우선 품사 태깅의 결과를 바탕으로 입력 어절의 각 자소에 품사를 할당하고, 형태소 분석 결과에서 몇 번째 형태소에 있는지를 적는다.

예를 들면, 신을 신고 신고하기의 품사 할당 결과는 [표 3]과 같다.

[표 3]

	0	0	0	1	1	1						
신을	nc	nc	nc	po	po	po						
	ㅅ	ㅣ	ㄴ	ㅇ	ㅡ	ㄹ						
	0	0	0	1	1	1						
신고	vb	vb	vb	ec	ec	ec						
	ㅅ	ㅣ	ㄴ	ㄱ	ㅗ							
	0	0	0	0	0	0	1	1	1	2	2	2
신고하기	na	na	na	na	na	na	xv	xv	xv	en	en	en
	ㅅ	ㅣ	ㄴ	ㄱ	ㅗ		ㅎ	ㅏ		ㄱ	ㅣ	

한편, 한국어의 표준 발음법은 중성 'ㅖ', 'ㅒ', ㅢ와 종성에 의해 발음 규칙이 일으켜지므로, 3개의 중성에 의한 발음 규칙과 27개의 종성에 의한 발음 규칙을 각각 작성하여 입력 어절의 중성과 종성을 보며 해당 규칙을 실행시킨다.

예를 들어 종성 'ㄴ'에 의한 규칙은 하기의 [표 4]와 같이 경음화, 자음 동화, 연음법칙 현상이 일어날 수 있는데, 모두 다음에 오는 초성의 품사에 의존하여 발생한다.

[표 4]

경음화 : 어간 받침 'ㄴ' 뒤에 결합되는 어미의 첫소리 'ㄱ, ㄷ, ㅅ, ㅈ'은 된소리로 발음한다.

자음동화 : 'ㄴ'은 'ㄹ'의 앞이나 뒤에서 [ㄹ]로 발음한다.

연음법칙 : 홑받침이나 쌍받침이 모음으로 시작된 조사나 어미, 접미사와 결합되는 경우에는 제 음가대로 뒤 음절 첫소리로 옮겨 발음한다.

그러므로 위의 예문의 경우, 전자의 '신고'는 경음화 현상이 발생되어 '신꼬'로, 후자의 '신고하기'는 어떠한 현상도 발생되지 않아 '신고하기'로 발음된다.

전술한 바와 같은 처리를 통해 추출된 형태소 분석열의 품사 정보가 발음 표기 변환 모듈(40)에 인가되면 발음 표기 모듈(40)은 대부분의 발음 변환 현상을 규칙에 의하여 처리하고, 합성어에서의 경음화 현상(예: 문고리[문꼬리]), 소리의 첨가 현상(예: 송이불[송니불]), 한자어에서의 경음화 현상(예: 갈증[갈쯩]에 해당하는 단어를 예외 발음 단어로 규정한다.

이렇게 품사 정보를 이용하여 입력 문장의 발음을 얻게 되지만, 어절의 정확한 발음을 얻기 위해서는 의미 처리 단계가 필요하다.

따라서, 입력 문장의 구문 구조를 알아내는 구문 분석기(50)는 확률적 의존 문법을 이용하여 입력 문장의 구조를 찾는다.

의존 문법은 구구조 문법과 달리 문장을 이루는 단어들의 지배소-의존소 관계를 찾기 위해 사용되고, 이는 일반적인 구구조 문법으로 표현 가능하다.

그러므로 구구조 문법을 이용하는 파싱 기법을 그대로 사용할 수 있다.

구문 분석기(50)는 구문 분석을 할 때, 어절들을 대표할 수 있는 비단말 기호(nonterminal symbol)가 필요하기 때문에 우선 품사 태깅된 결과를 바탕으로 각 어절들의 어절 품사를 생성한다.

어절 품사를 생성하는 방법은 입력 어절의 가장 왼쪽에 위치하는 형태소 품사와 가장 오른쪽에 위치하는 형태소 품사를 연결하여 만드는 것을 기본으로 하고, 만약 오른쪽에 쉼표 등 문장 기호들이 있을 경우에는 그 기호들을 더한다.

예를 들어 피어 있습니다.라는 어절은 피/vb+어/ex+있/vx+습니다/ef+./se로 형태소 분석이 되고, 어절 품사는 vbefse가 된다.

한편, 두 개의 형태소 품사가 더해져서 다른 형태소 품사로 변하는 경우가 있는데, 예를 들어 공부하다.의 경우 공부/na+하/xv+다/ef+./se로 형태소 분석이 되지만, '공부하'가 의미적으로 동사의 역할을 하므로 어절 품사를 vbefse 로 만든다.

이렇게 두 개의 형태소 품사가 하나의 형태소 품사로 바뀌는 규칙들은 다음과 같다.

1. 상태성 보통명사 + 형용사 파생 접미사 → 형용사

(예: 건강/ns+하/xj+다/ef+./se → vjefse)

2. 동작성 보통명사 + 동사 파생 접미사 → 동사

(예: 공부/na+하/vx+다/ef+./se → vbefse)

3. 보통 명사 + 명사 접미사 → 보통 명사

(예: 사람/nc+들/xn → nc)

4. 상태성 보통명사 + 부사 파생 접미사 → 부사

(예: 간단/ns+히/xa → ad)

5. 동사 + 부상 파생 접미사 → 부사

(예: 소중/ns+하/xj+게/xa → ad)

이상과 같이 문장을 이루는 어절들에 대해 어절 품사를 모두 얻은 후, CYK 테이블을 이용하여 확률적으

로 최적인 구문 구조를 찾는다.

일반적으로 의존 트리는 구구조 트리에 의해 표현될 수 있는데, 특히 한국어는 지배소 후위의 원칙이라는 특징이 있으므로 더욱 단순한 형태의 구구조 트리로 표현할 수 있다.

예를 들어 신을 신고 신고한다.라는 문장의 의존 트리는 첨부된 제7도이고, 이에 해당하는 구구조 트리는 첨부된 제8도와 같이 된다.

이 때 제8도의 구구조 트리에서 사용한 규칙들은 모두 다음의 세가지 형태 중 하나에 해당한다는 특징을 관찰할 수 있다.

1. S → A

(S : 시작 심벌, A: 어절 품사)

2. A → B A

(A, B : 어절 품사)

3. A → a

(A : 어절 품사, a : 어절)

한편, 제8도와 같은 구문 트리를 얻게 될 확률은 구문 트리에 사용된 모든 규칙들의 규칙 확률들의 곱으로 표현이 되는데, 규현된 구문 분석기에서는 첫 번째 규칙 형태와 세 번째 규칙 형태에 대해서는 모두 1.0의 확률을 부여하여 사용한다.

이는 문장의 마지막 어절이 항상 그 문장의 지배소가 되므로 최적 구문 트리를 찾는데 영향을 끼치지 않는다는 점과, 또한 어절 품사에서 임의의 어절이 발생할 확률은 품사 태깅의 최적 결과만을 이용할 경우, 최적 구문 트리를 찾는데 역시 영향을 끼치지 않는다는 점을 이용한 것이다.

그러므로 첨부된 제8도의 구문 트리 확률은 P(vbefse → vbec vbefse) * P(vbec → ncpo vbec)이 된다.

전술한 바와 같은 기능으로 실행되는 문서 분석기에서 품사 태거 성능 평가 실험 결과는 다음과 같다.

품사 태거(30)의 성능을 평가하기 위해 수동으로 태깅된 49,506어절의 학습 말뭉치로 모델을 학습시키고, 4,729어절에 대하여 실험하였다.

49,506어절 중 3,652어절은 미등록어에 관한 확률 $P(w_i|t_i)$을 학습시키기 위한 것으로, 나머지 45,854어절을 기반으로 사전을 구축한 후, 3,652어절에 대해 사전에 등록되지 않은 단어가 발견되면 이를 미등록어로 간주한다.

이때, 품사 태거의 성능은 하기의 [표 5]와 같고, [표 5]에서 등록 어절과 미등록 어절이란 각각 미등록어가 포함안된 어절과 미등록어가 포함된 어절을 뜻한다.

[표 5]

		어절단위	형태소 단위
전체어절	4729 어절 9980 형태소	85.11%	90.33%
등록어절	4196 어절 8869 형태소	88.03%	93.46%
미등록어절	533 어절 1111 형태소	62.10%	65.25%

전술한 바와 같이 [표 5]에서 등록 어절에 대한 정확률이 88.03%로 그다지 높지 않은 정확률을 나타내는 것을 관찰할 수 있는데, 이는 구현된 태거가 형태소 격자가 구성되지 않았을 경우에만 그 어절에 미등록어가 포함되었다고 가정하는 것에 기인한다.

예를 들어, 신을 신고 신고한다라는 문장에서 '신을'의 올바른 태깅 결과는 '신/보통명사+을/목적격조사'인데 만약 '신'에 대해 동사라는 정보만 사전에 있을 경우, 형태소 분석기는 '신/동사+ㄹ/관형사형전성어미'의 형태소 격자만을 작성하고, 품사 태거는 미등록어가 없다고 간주한 우에 이를 그대로 태깅하게 된다.

바로 이런 종류의 오류 때문에 미등록어가 포함안된 어절에 대한 태깅 결과가 그다지 높지 않게 되었다.

한편, 실험 말뭉치에서 사용되는 단어들을 사전에 첨가한 후 태깅한 결과, 어절에 대한 정확률이 96.46%, 형태소에 대해서는 98.01%의 정확률을 얻었다.

이러한 실험 결과를 볼 때, 사전의 표제어 수가 높아질수록 태거의 성능은 더 높아진다.구문 분석기(50)의 성능 평가 실험결과는 다음과 같다.

수동으로 작성된 498문장의 구문 트리들을 바탕으로 모델을 학습시킨 후, 100문장에 대해 실험하였다.

100문장은 738어절로 이루어졌고 문장의 마지막 어절은 항상 자신을 지배소로 결정하므로 638개의 지배소-의존소 관계가 존재하게 된다.

우선 품사 태거의 결과를 모두 맞게 만들어 구문 분석기 자체의 성능을 평가하는 실험과 품사 태거의 결과를 그대로 이용하는 실험을 하였는데, 전자의 경우 80.87%의 정확률을 나타내었고, 후자의 경우 78.68%의 정확률을 나타내었다.

이 결과는 예를 들어 약 11개의 어절로 된 문장에 대해 마지막 어절을 제외한 어절 중 8개의 어절은 자신의 지배소를 올바르게 선택한 것으로 볼 수 있다.

발명의 효과

이상에서 설명한 바와 같이, 본 발명에 따른 문서 분석기는 입력 문서에 대해 비결정 유한 오토마타를 이용하여 전처리를 수행한 후, 형태소 분석 단계에서 어절의 모든 가능한 형태소 분석 결과를 구하며, 형태소 분석 결과들 중 확률적으로 가장 가능한 형태소 분석 결과를 선택한 후, 선택된 결과를 발음 표기 변환 모듈을 통해 어절을 이루는 모든 자소에 품사를 할당한 후, 3개의 중성에 의한 규칙, 27개의 종성에 의한 규칙들을 적용하여 어절의 발음 표기를 얻은 다음 구문 분석기의 확률 의존 문법을 이용하여 입력 어절들의 지배소-의존소 관계를 찾아 운율을 생성하여 신뢰성 있는 문서 음성으로 변환을 제공한다.

본 발명에 따른 문서 분석기는 품사 태거에서 85.11%의 정확률을, 구문 분석기에서 78.68%의 정확률을 각각 나타내며, 특히 품사 태거의 경우, 처리 중인 문서에 미등록어가 없을 경우에는 96.46%의 높은 정확률을 제공한다.

또한, 본 발명에 따른 문서 분석기는 문서 음성 변환 시시템 이외에도 음성 인식 시스템을 위한 발음 사전을 구축하고자 할 때 유용하게 사용될 수 있다.

(57) 청구의 범위

청구항 1

문서 음성 변환 시스템에 있어서, 음성 변환을 위하여 입력되는 문서를 하나의 문장씩 추출하며, 비결정 유한 오토마타를 이용하여 비한글 문자들을 한글 문자로 변환시키는 전처리 수단과; 상기 전처리 수단에서 인가되는 하나씩의 문장 어절에서 모든 가능한 형태소 분석 결과를 구하는 형태소 분석수단과; 상기 형태소 분석된 입력 문장을 확률 정보를 기반으로 하여 가장 가능성이 높은 형태소 분석 결과를 선택하여 최적 형태소 분석열의 결과를 이용하여 각 어절들의 발음 표기를 구하는 발음 표기 변환수단 및; 확률 의존 문접을 이용하여 어절들의 지배소-의존소 관계를 구하며, 입력된 문서에 대한 어절의 발음 표기 및 형태소 품사열, 의존 트리를 최종 결과로 출력하는 구문 분석수단을 구비하는 것을 특징으로 하는 한국어 문서 음성 변환 시스템을 위한 문서 분석기.

청구항 2

청구항1에 있어서, 상기 품사 태거와 구문 분석수단에는 중의성 해소를 수행하기 위해 각각 품사 태깅을 확률 정보와 구문 분석을 위한 확률 정보를 구비하는 것을 특징으로 하는 한국어 문서 음성 변환 시스템을 위한 문서 분석기.

청구항 3

청구항1에 있어서, 상기 전처리 수단은 비한글 문자들에 대하여 영어, 영어, 약어, 숫자, 전화번호, 연도, 시간 등으로 구분하며, 이들을 다음에 오는 한글 문자들을 고려하여 한글화시키는 것을 특징으로 하는 한국어 문서 음성 변환 시스템을 위한 문서 분석기.

청구항 4

청구항1에 있어서, 상기 전처리 수단은 총 48개의 상태로 이루어지는 비결정 유한 오토마타로 구현되며, 이중 12개는 종결 상태가 각각 다른 비한글의 한글 변환 모듈로 이루어지는 것을 특징으로 하는 한국어 문서 음성 변환 시스템을 위한 문서 분석기.

청구항 5

청구항1에 있어서, 상기 형태소 분석수단은 입력된 문서에서 판독된 단어를 사전에서 참조할 때 예외 발음 단어일 경우 발음 표기를 발음 표기 변환수단에 직접, 넘겨주는 것을 특징으로 하는 한국어 문서 음성 변환 시스템을 위한 문서 분석기.

청구항 6

청구항1에 있어서, 상기 형태소 분석수단은 전처리 수단에서 인가되는 문장에 대하여 형태소 분석을 실패한 경우 분석되는 어절에 미등록어가 있다고 판단하고 미등록어를 추정한 후 언어적 휴리스틱을 이용하여 그 후보를 최소화시키는 것을 특징으로 하는 한국어 문서 음성 변환 시스템을 위한 문서 분석기.

청구항 7

청구항1에 있어서, 상기 형태소 분석수단에 설정되는 격자 구조 알고리즘을 다음과 같이 이루어지는 것을 특징으로 하는 한국어 문서 음성 변환 시스템을 위한 문서 분석기.

청구항 8

청구항1에 있어서, 상기 형태소 분석수단은 입력 어절의 불규칙 및 축약 현상 위치들을 미리 계산한 다음 격자 구성 알고리즘에 따라 형태소 분석하여 분석된 결과를 형태소 격자로 표현하는 것을 특징으로 하는 한국어 문서 음성 변환 시스템을 위한 문서 분석기.

청구항 9

청구항1에 있어서, 상기 형태소 분석수단은 입력되는 문장 어절의 오른쪽에서 왼쪽으로 형태소를 찾아 형태소 격자를 구성하는 것을 특징으로 하는 한국어 문서 음성 변환 시스템을 위한 문서 분석기.

청구항 10

청구항1에 있어서, 상기 형태소 분석수단은 음절정보를 사용하여 미등록어의 마지막 음절이 추정된 음절로 사용되지 않는 경우 그 노드를 격자에서 제외하는 것을 특징으로 하는 한국어 문서 음성 변환 시스템을 위한 문서 분석기.

청구항 11

청구항1에 있어서, 상기 형태소 분석수단은 단서 형태소를 이용하여 미완성된 형태소 격자내에 아주 빈도가 높고 그 어절의 구성을 추정하기에 충분하다고 생각되는 형태소가 발견되면 그 형태소의 앞에 추정된 미등록어만 남기고 나머지를 격자에서 제거하는 것을 특징으로 하는 한국어 문서 음성 변환 시스템을 위한 문서 분석기.

청구항 12

청구항1에 있어서, 상기 품사 태거에서 실행되는 품사 태거 수식은 다음과 같이 이루어지는 것을 특징으로 하는 한국어 문서 음성 변환 시스템을 위한 문서 분석기.

$$\phi(w_{1..n}) = \arg\max_{m_{1..n},t_{1..n}} \prod_{i=1}^{n} [P(m_i|t_i) \prod_{j=2}^{N_i} P(t_i^j|t_i^{j-1})] P(t_i^1|t_{i-1}^{N_{i-1}})$$

$$P(m_i|t_i) \cong P(k_i^1|t_i^1)[k_i^1 \prod_{j=1}^{N_i} P(m_i^j|t_i^j) +$$

$$(1-k_i^1)P(t_i^1|t_i^2 m_i^2) \prod_{j=2}^{N_i} P(m_i^j|t_i^j)]$$

청구항 13

청구항1에 있어서, 상기 발음 표기 변환수단은 어절을 이루는 모든 자소에 품사를 할당한 다음 3개의 중성에 의한 규칙과 27개의 종성에 의한 규칙으로 어절의 발음 표기를 얻는 것을 특징으로 하는 한국어 문서 음성 변환 시스템을 위한 문서 분석기.

청구항 14

청구항1에 있어서, 상기 발음 표기 변환수단은 합성에서의 경음화 현상과 소리의 첨가 현상, 한자어에서의 경음화 현상에 해당하는 단어를 예외 발음 단어로 규정하는 것을 특징으로 하는 한국어 문서 음성 변환 시스템을 위한 문서 분석기.

청구항 15

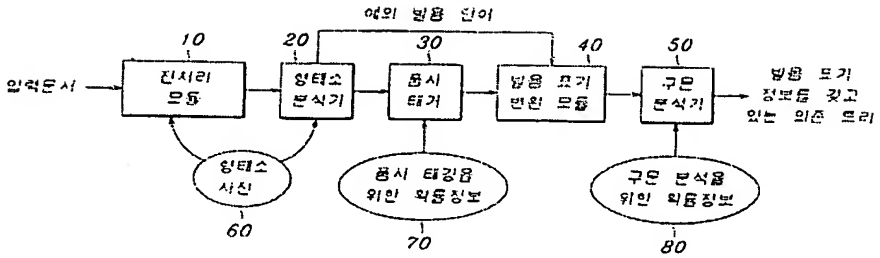
청구항1에 있어서, 상기 구문 분석수단은 입력 어절의 가장 왼쪽에 위치하는 형태소 품사와 가장 오른쪽에 위치하는 형태소 품사를 연결하여 기본 적인 어절 품사를 생성하는 것을 특징으로 하는 한국어 문서 음성 변환 시스템을 위한 문서 분석기.

청구항 16

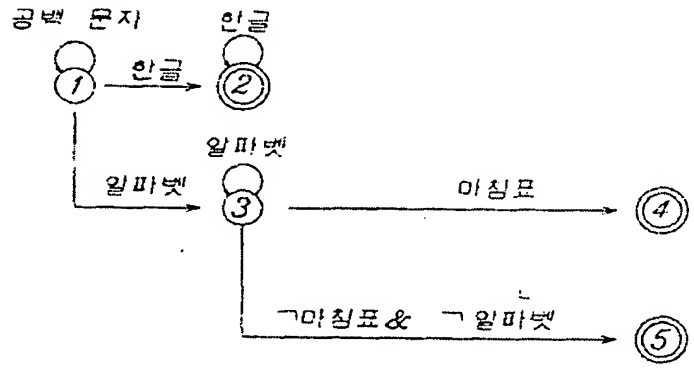
청구항1에 있어서, 상기 구문 오른쪽에 쉼표 등 문장부호가 있는 경우에는 그 기호를 더하여 어절 품사를 생성하는 것을 특징으로 하는 한국어 문서 음성 변환 시스템을 위한 문서 분석기.

도면

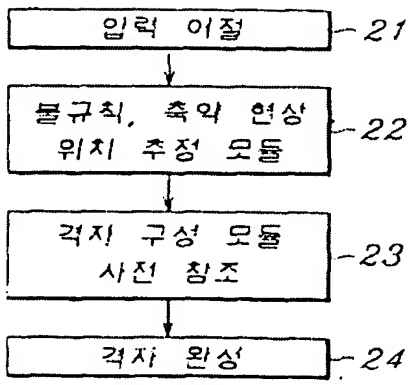
도면1



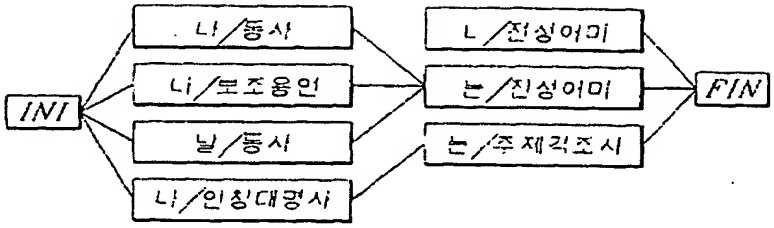
도면2



도면3



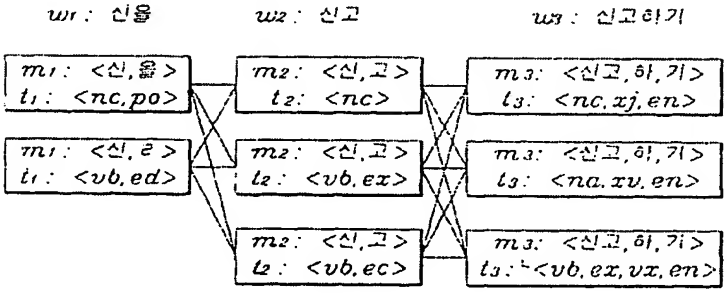
도면4



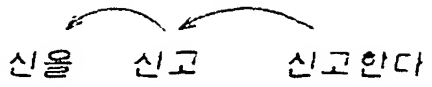
도면5

```
i ← 마지막 자소의 오른쪽 번호
k ← 0
J ← {i}
L ← {(k, FIN, NULL,{ })}
while i > 0 do
  i ← i - step
    for all j ∈ J do
       T ← LookupDict(w_ij )
       for all t ∈ T do
           k ← k + 1
           I ← SearchL((w_ij , t))
           P ← { }
           if I is not empty then
               L ← L ∪ {(k, w_ij , t, I)}
               P ← {i}
           end if
       end for
    end for
    J ← J ∪ P
    불규칙 및 탈락 현상을 처리
end while
k ← k + 1
I ← SearchL((INI, NULL))
L ← L ∪ {(k, INI, NULL, I)}
```

도면6



도면7



도면8